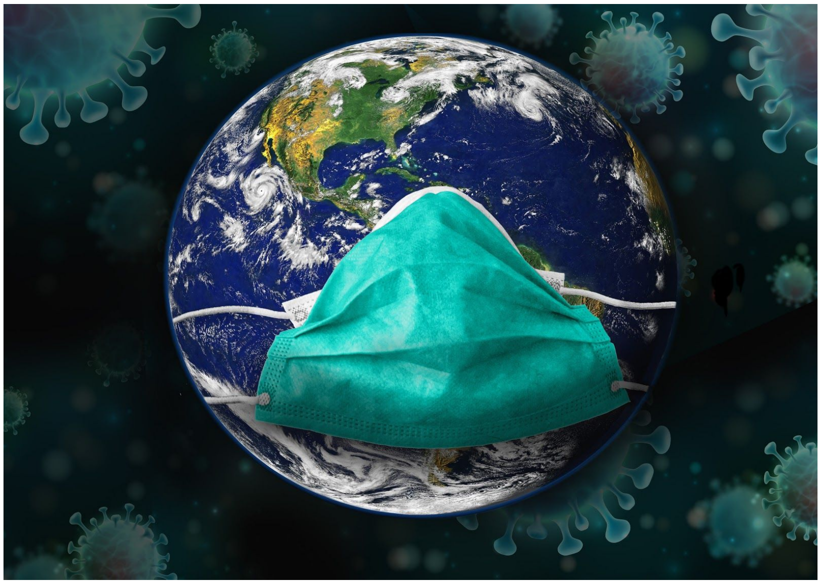


# COVID-19: Impacts and Insights

Manar Hasan, Pairode Jaroensri, Kshitij Korde, Maniratnam Mandal, Akib Shah

### *Abstract*

The COVID-19 pandemic has affected our way of life for most of this year. Researchers all over the world have utilized machine learning techniques to analyze the pandemic and its effects. This project was divided into three separate analytical tasks - measure the efficacy of government interventions, model current projections, and analyze the pandemic's effect on mental health. Imperial College London's model was replicated for the first task, with visualizations for various regions clearly showing the effect of government interventions on cases and deaths. Current projections were modelled using standard statistical models and learning-based models, with Stacked LSTM showing comparatively best results. Mental health analysis was modelled as a classification problem, with different models performing with varying degrees of success depending on how the data was preprocessed.

### *Introduction and Background*

The COVID-19 pandemic has raged through the entire world, affecting our way of life for most of this year. More than 68 million people have been infected until now, with a death toll of over 1.5 million and rising [1]. Only a couple of months ago, when this project was proposed, those numbers were 42 million infections and 1.1 million deaths. Although there has been no definite treatment developed yet (and vaccines have only just been developed and will require months, if not years, to fully propagate among the world population), countries have implemented several preventive measures to halt the spread. We observe that countries like New Zealand or Vietnam have been successful in bringing cases down to almost zero [2, 3]. One major concern with the pandemic is the spread of misinformation and subsequent denial about the severity (or even existence) of the disease. On the positive side, because our lives and policies are so data-driven, enormous and diverse data have been collected pertaining to different aspects of COVID-19 [4]. Utilizing these data sources, researchers have developed models to predict and analyze the impacts of the pandemic [5]. The goals for this project are to use our knowledge in data mining and machine learning to leverage the data to - gain insights from it through elaborate visualization, analyze the impact of preventative policies, predict infected projections, and compare results for different scenarios.

Because of the wealth of data available, this project will not be the first (or the last) to leverage data to understand the pandemic at a deeper level. Researchers all over the world have utilized machine learning techniques to analyze the pandemic and its effects - from predicting cases, predicting mortality, identifying at-risk populations, to even predicting the next pandemic [6, 7, 8]. This project aims to add to that body of knowledge and understanding. More specifically, this project will be divided into three separate tasks, described as below:

- Task 1. Efficacy of non-pharmaceutical interventions - To measure how effective various preventive measures implemented by various world governments were (such as social distancing or mandating face-coverings), this task aims to compare and contrast spread trends before and after these interventions. These analyses will be done within countries, as well as comparing between countries that vary in levels of proactivity.

- Task 2. Projection - This task aims to model the current projection of infections and deaths using time-series statistics, by replicating other oft-used prediction models.

- Task 3. Effect on mental well-being - Lastly, this task aims to gain insights regarding the effect of COVID-19 on the well-being of people. Specifically, US census data will be consulted to infer about how COVID-19 impacted the people's mental health.

This project aims to contribute to the growing pool of knowledge in a few different ways. By highlighting the effectiveness of common preventative measures, this work can support decision-making by world leaders, showing clearly the effects of various measures in different countries. By calibrating and replicating predictive models, this work can corroborate the work done by others, increasing the credibility of such work. By analyzing mental health effects of the pandemic, this work can be consulted by community leaders looking for data-driven methods to alleviate the negative mental health impacts of the pandemic, as well as allocating funding more appropriately.

The following sections of this blog post are organized as follows - each task is laid out one by one, with a description of each task’s data, the model(s) used, assumptions made in the analysis, results achieved, and associated challenges with the task. After all of the above is described for all three tasks, overall conclusions close out this project blog post.

# Task 1: Efficacy of non-pharmaceutical interventions

### *Introduction and Background*

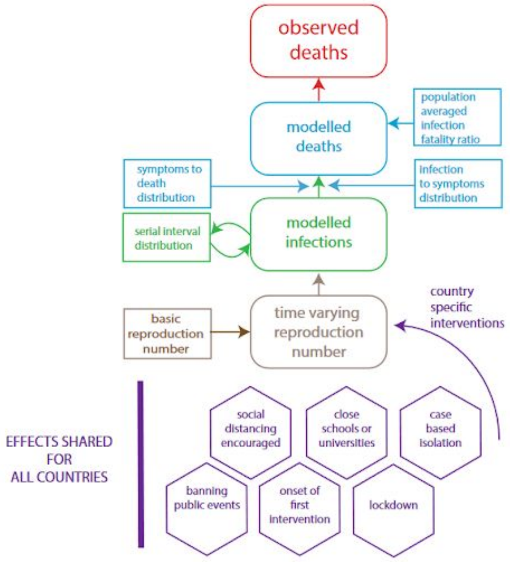
Since the beginning of the pandemic, researchers across the world have modelled the projection using sophisticated statistical methods to aid the administrative policymaking and implementation. There have been various competing models often with different projection results [CDC]. As intervention policies were being issued upon recommendations of WHO, they were also being widely debated. Hence, it was and is still very important to analyze the case and death data [1], and determine the efficacy of the government interventions to plan ahead. The available data shows how the effect of the interventions can affect different demographics in different countries, which can be effectively utilized for controlling the pandemic now that we are observing a resurgence of cases globally.

While searching for models to study the efficacy, we looked for models with transparency and reproducibility. Upon careful research, we chose the Covid19Model by the Imperial College London team, which specifically focuses on the impact of non-pharmaceutical interventions in 11 European countries (which was later augmented to Brazil and USA). The first task of this project was thus reproducing and studying the validity of results produced by this model.

### *Data Description*

The daily consolidated confirmed cases and deaths were obtained from the ECDC for 11 European countries. The age-stratified population data were obtained from the United Nations Population Division [report]. Nature and type of non-pharmaceutical interventions, along with their date of implementation, were collected from government and corresponding public health webpages. Because the aim was to study the effect of interventions, and as most of the countries had them in place during the first wave of cases, the model was trained on the data till the first rise and fall of cases/deaths.

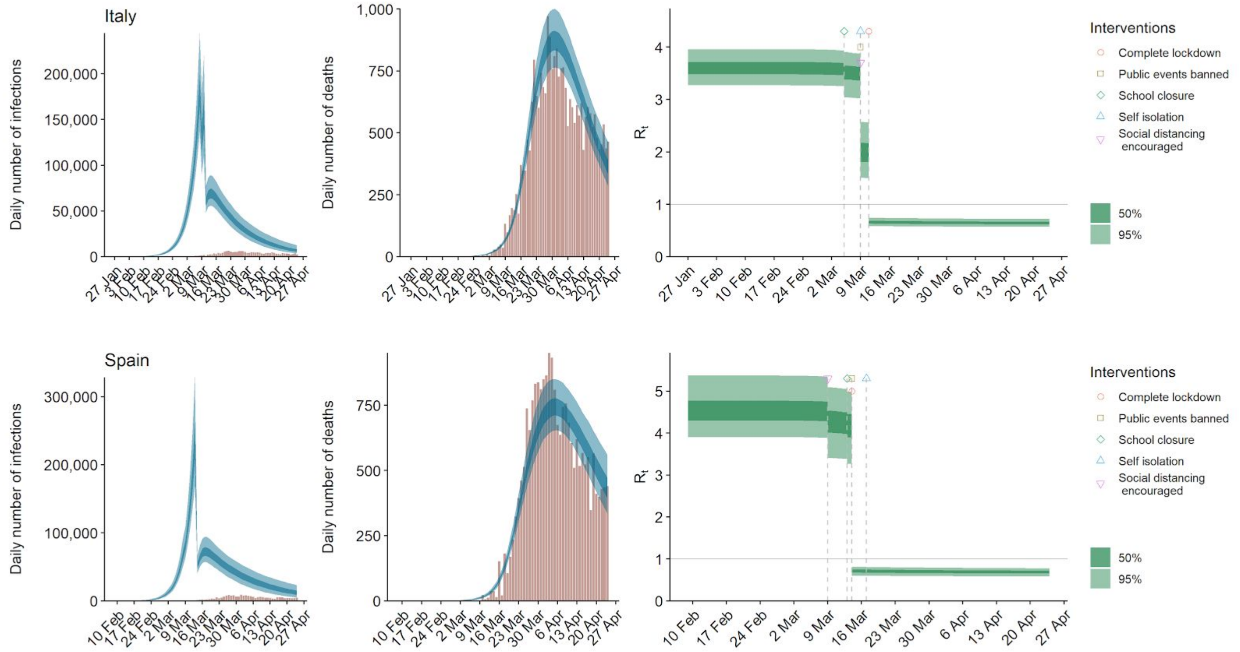
### *Modelling*


observed deaths
modelled deaths
population averaged infection fatality ratio
symptoms to death distribution
infection to symptoms distribution
serial interval distribution
modelled infections
basic reproduction number
time varying reproduction number
country specific interventions
EFFECTS SHARED FOR ALL COUNTRIES
social distancing encouraged
close schools or universities
case based isolation
banning public events
onset of first intervention
lockdown

The model is based on transmissibility, and the Bayesian hierarchy of the framework from the modelled infections to deaths is as shown in the figure (left). The number of deaths is modelled based on the infection-to-death distribution, population-averaged fatality ratio, and modelled infections. Whereas, the number of infections is modelled as a function of the reproduction number ($R_t$) and serial interval distribution (distribution of the time it takes for an infected to infect another). $R_t$ is a fundamental parameter used to study the spread of diseases and describes the number of people an infected person transmits the disease to at time *t*. Essentially, it reflects the exponential nature of the spread, and a value less than 1 will indicate that the spread is slowing down. This reproduction number is modelled to be influenced by the interventions. So, to summarize, the model estimates the number of deaths, the number of cases, and the reproduction number for each country. For implementation, it is jointly fit for all 11 countries, with both individual and shared effects on $R_t$. It is cross-validated over a 14-days.

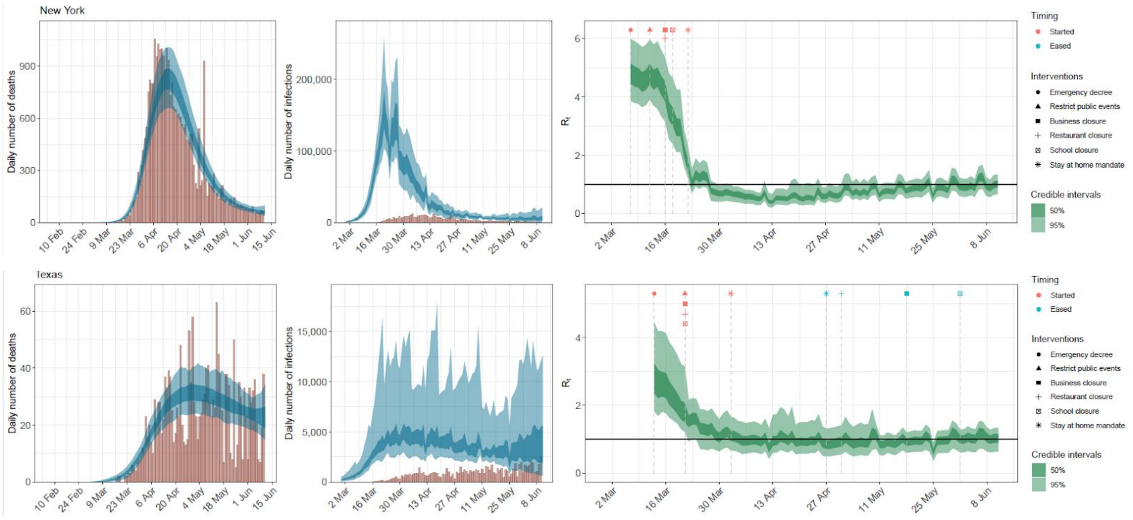
***Results***

Let's focus on two countries which were the centres of attention during the initial pandemic crisis - Italy and Spain.



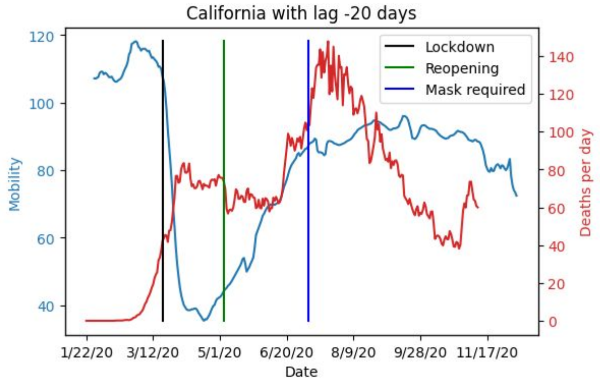
The left plot shows the actual number of infections estimated, which is a lot more than the recorded number, as expected. The estimated death projection (middle) closely follows the recorded number. Modelling based on recorded deaths is more desirable as it is evidently more accurate to the real deaths accountable to COVID-19 than the number of recorded cases is to the original number. The right plot shows the effect the interventions had on the Rt, which dropped from a high value to a number less than 1, which corroborates the decreasing trend of the number of deaths or that the pandemic is slowing down. It was also assumed that the change in $R_t$ was immediate which is more gradual in reality.

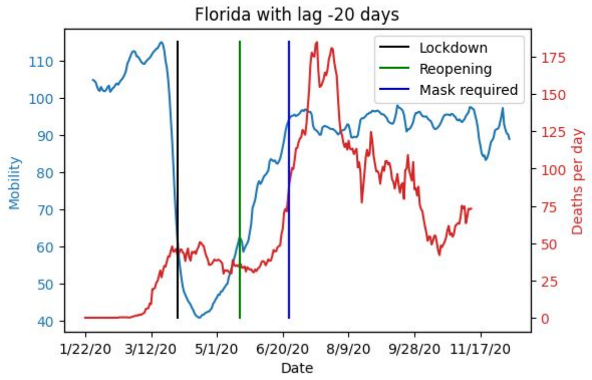
The later version of the model was also implemented for Brazil and the US, this time also including mobility data as a parameter. The plots for New York and Texas are shown below.

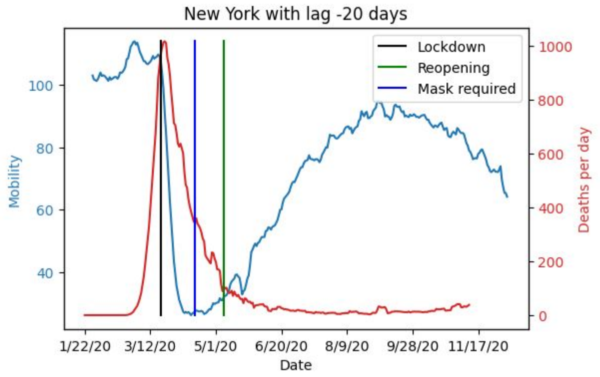


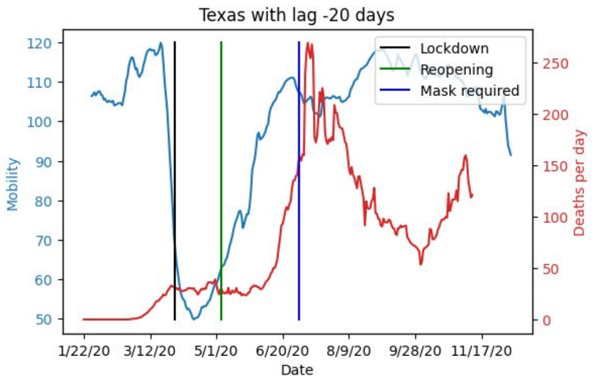
### *Visualization*

To gain some general understanding of the current COVID-19 situation, Let's look at a few visualizations. We plot the number of new deaths per day v.s. the Apple Mobility Index for comparison. According to IHME, COVID-19 takes about 20 days median from infection to death, so we shift the death plot forward by 20 days in an attempt to line it up with the date of actual infection. The first set of plots is for California, New York, Florida, and Texas, and we also mark the dates each state implemented lockdown, face mask mandate, and beginning of reopening.



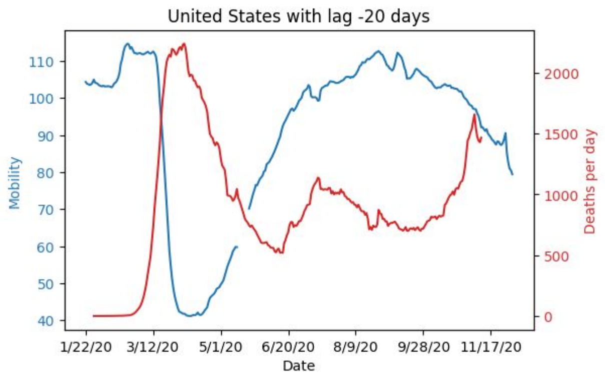


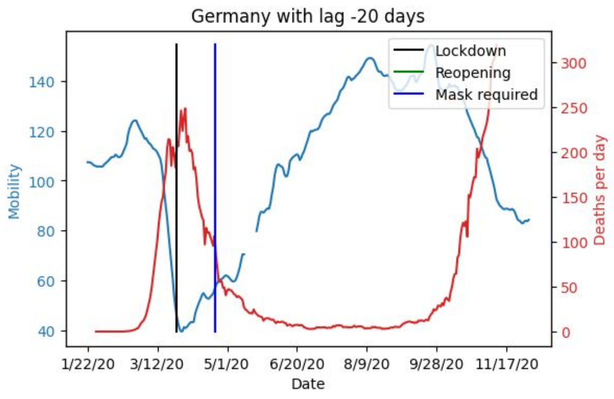


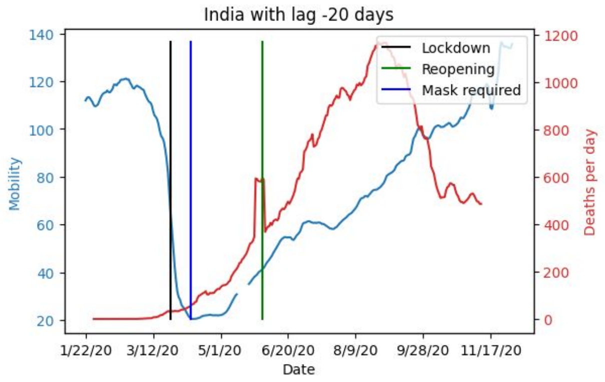


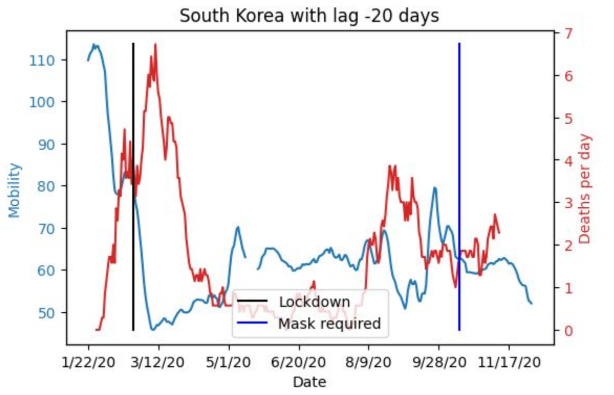
For all 4 states, we can see that after quarantine, deaths per day either stops increasing or even decreases. When the states begin reopening, 3 states see increases in deaths per day. New York, however, does not see an increase in deaths per day after reopening.

Next, we inspect the *mobility* v.s. *deaths per day* plot for Germany, South Korea, and India compared to the US as a whole. Germany, South Korea, and the US follow the same trend, in which deaths per day decrease after the initial lockdown, then increases into a second spike after reopening. India, however, only has one increase of deaths per day into one big peak before decreasing again.

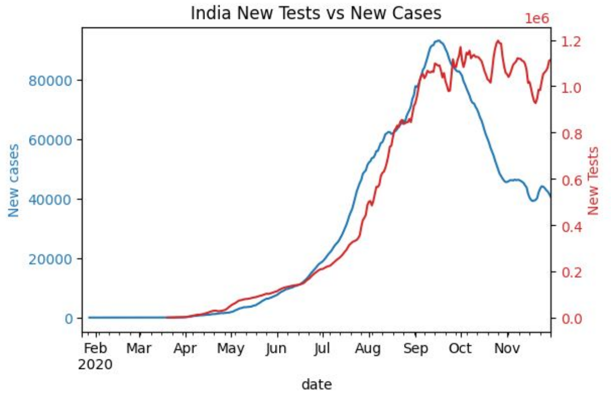


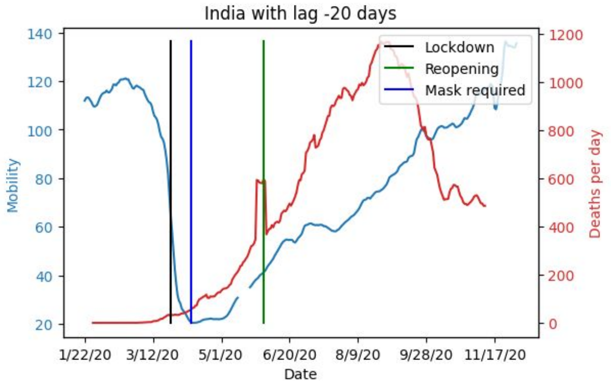




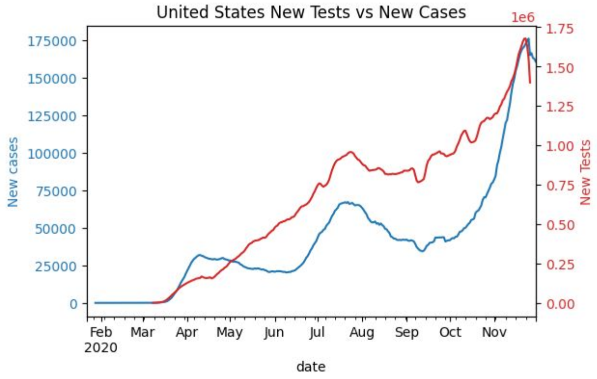


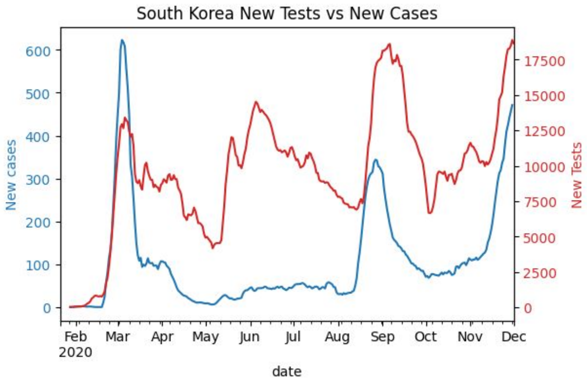
To investigate India further, we plot the number of *new tests per day* v.s. *the number of new confirmed cases per day*. From the plot, we can see that both curves follow the same increasing trend until the number of confirmed cases reaches its peak before dropping down, but the number of new tests stays mostly the same. This may be an indicator that India did not have an adequate number of tests, so the number of new cases is capped by the number of new tests, which is why India's trend is different from other countries analyzed.


India New Tests vs New Cases


India with lag -20 days

Additionally, we have heard a claim from the news that the number of cases in the US seems like a lot only because we have a lot of tests. To investigate this claim, we plot the number of new cases v.s. the number of new tests for the US and South Korea. We can see that both India and the US do see a rising number of cases with more testing, so there might be some truth to the claim. South Korea however, shows an exception. The number of new cases in South Korea follows the trend of the number of new tests only for the first peak, after which, they don't seem to follow the same trend anymore.


United States New Tests vs New Cases


South Korea New Tests vs New Cases

Although there are a few exceptions, we can see that for the countries that we inspect, the number of deaths per day either decreases or stays level when the lockdown measures are implemented, and rises again after they begin reopening. The number of tests also affects the number of cases as more tests may reveal more cases. Lastly, we would like to remind everyone that although there might be a clear trend in our data, correlation is not causation. We can only establish causation by performing appropriate experimentation.

## Task 2: Projections using Time-Series analysis

### *Introduction and Background*

One of the most important aspects of pandemic control is projecting the number of infections and deaths based on the current data. Having failed to successfully reproduce and validate currently popular models, we approached the projection problem as basic time-series estimation. Time-series projection is a standard problem in data science and we tried our hands on both standard statistical models (for short-term projection) and advanced learning-based models (for long-term projection).

### *Data Description*

As we explore basic models, we only used the current case and death data as inputs and did not consider other parameters essential to epidemiological models. For daily count, we used the Johns Hopkins CSSEGIS US data, which is updated daily. For our project, we used data from the beginning (22nd Jan) till the end of November. The data for all states were accumulated and the target variable was nationwide counts. Also, we only considered the range from the date when the first death was recorded (29th Feb). Both daily and cumulative counts were considered as inputs, but the daily data was log-transformed to help stabilize the variance before modelling.

### *Modelling*

*Standard Statistical Models* - The goal of short term projection is to estimate the count of one day based on previously observed data. So, the estimates can be based on all of the previously observed data or a fixed time period. We tackled both, the daily case projection and daily death projection problems. The following basic and advanced time-series estimators were compared:

- Mean Constant Model
- Linear Trend Model
- Linear Model with Regressor
- Random Walk Model
- Simple Moving Average (SMA) over 7 days
- Simple Exponential Smoothing (SES) with half-life of 7 days
- Auto-Regressive Integrated Moving Average (ARIMA) model

We chose a seven-day window as we had observed weekly seasonality in the data upon additive decomposition. This is just because of the way deaths and cases are reported, and has nothing to do with how they actually occur. For the ARIMA model, the auto-regressive (AR) terms and moving average (MA) terms were determined using auto-correlation and partial auto-correlation functions respectively.

*Learning-based Models* - Long Short-Term Memory networks (LSTMs) have been used to model univariate time series forecasting because of their ability to learn patterns in the data. Also, LSTMs outputs can be extrapolated for long-term forecasting. So, we train the models using the data till the end of October and compare them on the basis of forecasting for the whole month of November. Cumulative case and death data were used for the respective problems, as we found them to give better results than the daily data. The various variations of LSTMs that have been compared are:

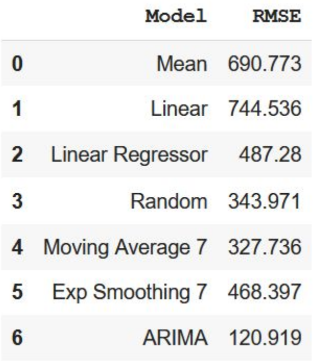
- Vanilla LSTM
- Stacked LSTM
- Bidirectional LSTM
- CNN LSTM
- ConvLSTM

For the first three variants, input sequence of length 7 (days) was used with 20 hidden layer units. For the last two variants, because of their convolutional nature, we used a six-day input sequence broken into two sub-sequences.

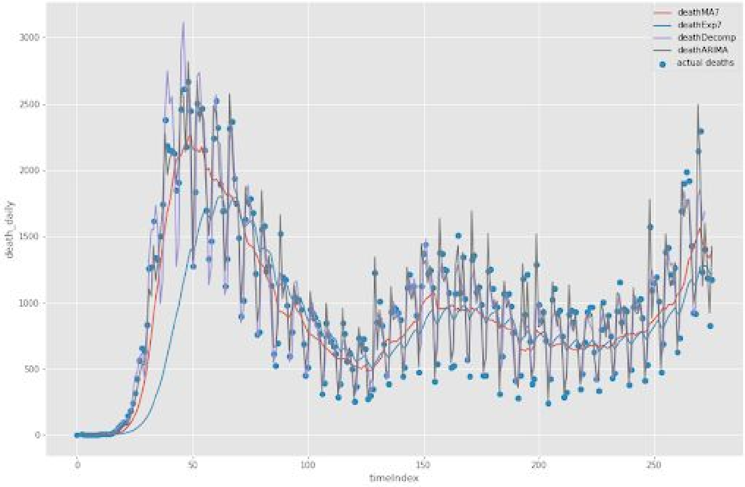
***Results***

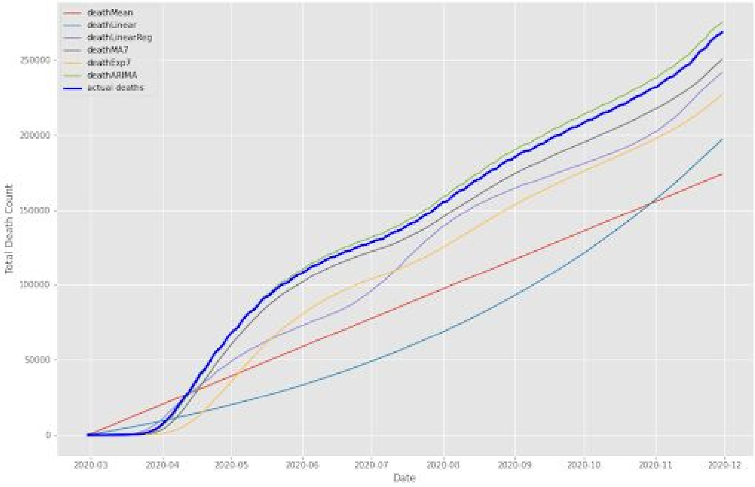
*Standard Models:*

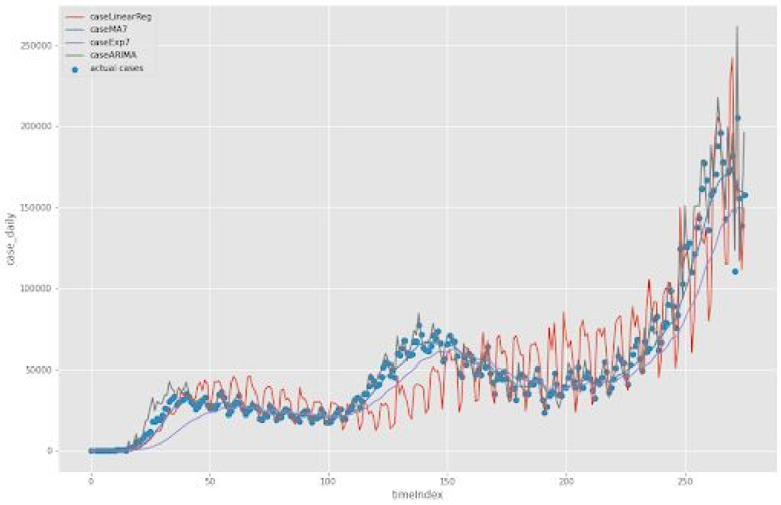
The RMSE values for all the models are shown on the left for the daily death estimates. The daily and cumulative projection plots are shown below. ARIMA model performs the best in both the cases, which is expected as it is the most advanced among the standard statistical models. Plots for daily and cumulative death counts and case counts are shown below.

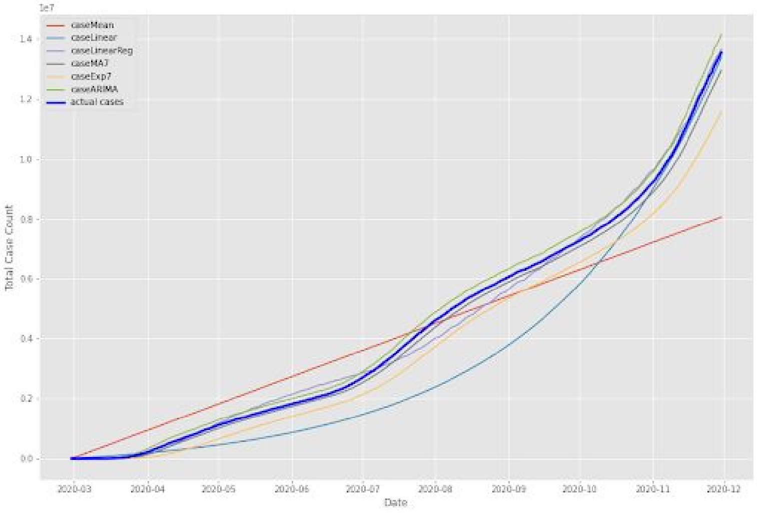


| | Model | RMSE |
|---|---|---|
| 0 | Mean | 690.773 |
| 1 | Linear | 744.536 |
| 2 | Linear Regressor | 487.28 |
| 3 | Random | 343.971 |
| 4 | Moving Average 7 | 327.736 |
| 5 | Exp Smoothing 7 | 468.397 |
| 6 | ARIMA | 120.919 |

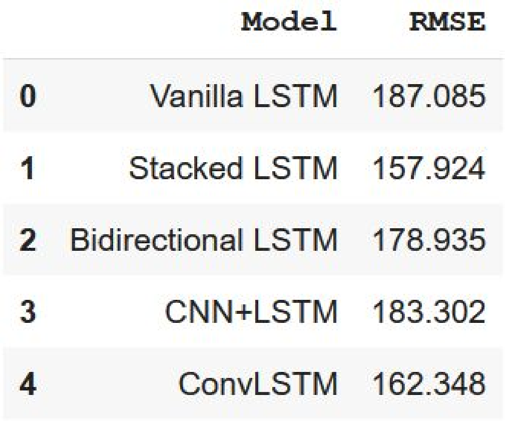




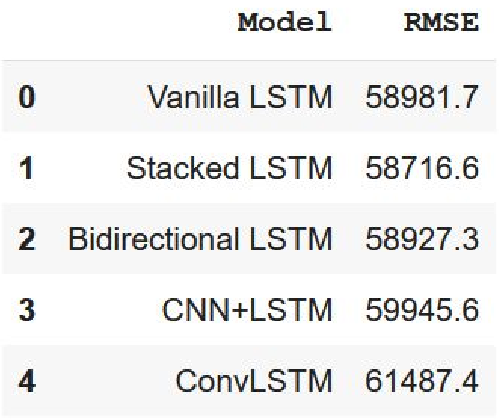




*LSTM-based Models:*
The cumulative and daily projection plots are shown for both the infection and death estimates. While the death estimates are very close to actual, the case projections are not so accurate. But both for both the cases, models were able to project an increasing trend. The RMSE of the estimates are also given. The Stacked LSTM model performed comparatively well for both cases. ConvLSTM was the most unstable during training, probably because of model complexity.

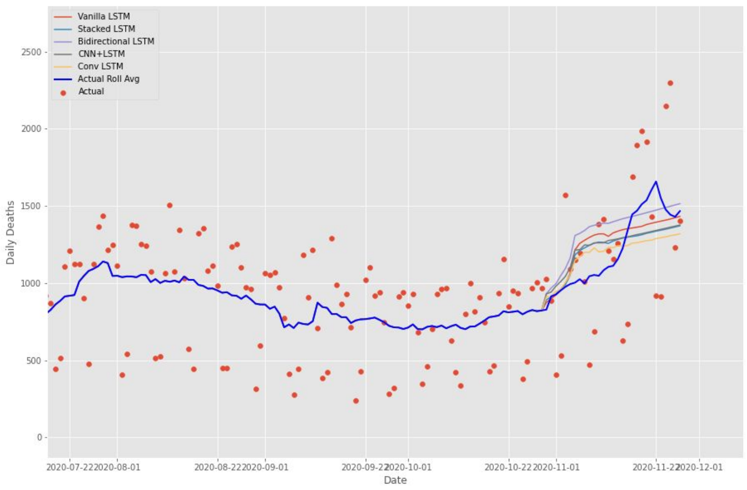


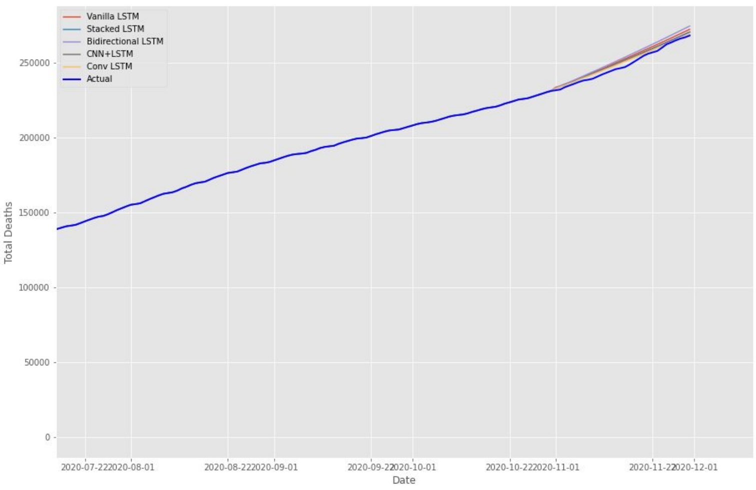
| | Model | RMSE |
|---|---|---|
| **0** | Vanilla LSTM | 187.085 |
| **1** | Stacked LSTM | 157.924 |
| **2** | Bidirectional LSTM | 178.935 |
| **3** | CNN+LSTM | 183.302 |
| **4** | ConvLSTM | 162.348 |

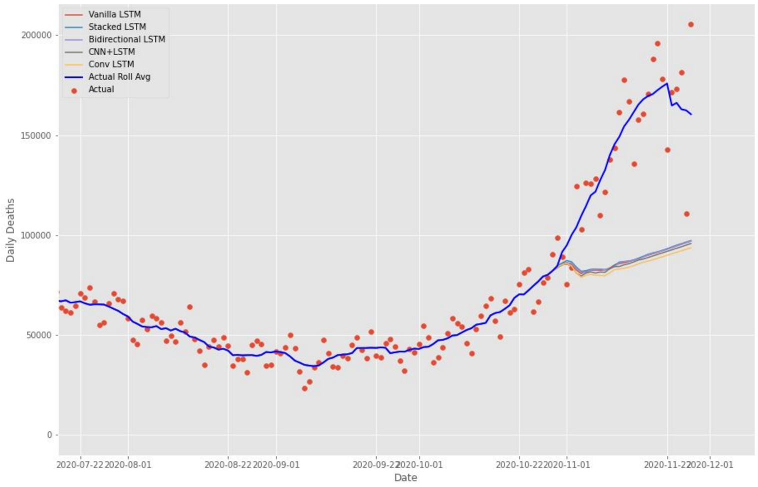


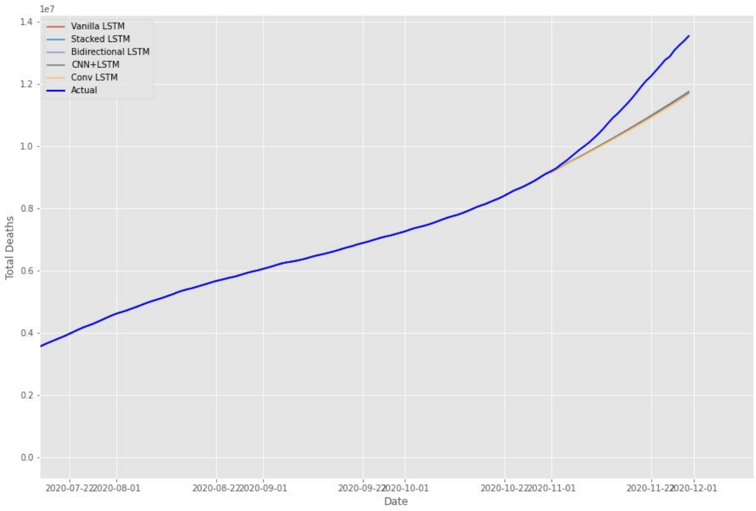
| | Model | RMSE |
|---|---|---|
| **0** | Vanilla LSTM | 58981.7 |
| **1** | Stacked LSTM | 58716.6 |
| **2** | Bidirectional LSTM | 58927.3 |
| **3** | CNN+LSTM | 59945.6 |
| **4** | ConvLSTM | 61487.4 |

The daily and cumulative death plots and case plots are shown in the next page. The recorded data is in blue. Plot range is from the beginning of August to the end of November. The projected data and calculated RMSE is for the month of November.









## Task 3: Effect on mental well-being

### *Introduction and Background*

The COVID-19 pandemic has affected people not just physically but also mentally. Many people are suffering from anxiety and depression as an effect of the pandemic. Our aim is to try to figure out which of the people can be suffering from mental health issues using demographic data. This way we can proactively reach out to people in the hope of helping those who might be affected.

Data mining algorithms are used to identify the various mental health issues [9]. People have started to understand the importance of pandemic and its impact on mental health [10, 11].

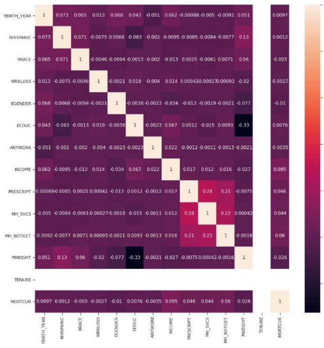
In order to achieve the aim, we are using the data from the surveys conducted by the US Census Bureau and train classifiers for predicting whether a person might be facing mental health issues (binary classification) and to what extent he/she might be affected(multi-class classification).

As students, we have always been taught the importance of mental health. It is hard to tell whether a person might be facing mental health issues. We took this opportunity to make a novel contribution to the field.

***Data Description and Exploration***

The United States Census Bureau has been collecting and disseminating Household Pulse Survey data on a weekly basis since April 23, 2020. Data for the nation, each of the fifty states, plus Washington, D.C., and the fifteen largest metropolitan areas are collected as responses from individuals. The data has responses for questions that cover Education, Employment, Food Sufficiency and Security, Health, Housing, Social Security, Spending and Transportation.

The primary dataset chosen for modelling was a set of responses collected between September 30th to October 12th (week 16) which is a part of the second phase of the survey. The total number of relevant columns in the set after dropping the details like the week of data collection, region and allocation flags come close to 160 in number. Each of these columns is a specific question to which an individual has responded in terms of their opinion on the question. While each row is the collection of responses for a single individual. The data is a mix of categorical and non-categorical values depending on the kind of question where the responses are in a specified range of values. The data also has missing and ignored responses represented by (-99) for questions seen but category not selected and (-88) Missing / Did not report. The questions with responses that could closely represent the mental health of the respondent are separated as the target for modelling. These include the questions checking if the individual felt anxious, worried, disinterested or depressed in the last 7 days. The responses were representative of the frequency of experiencing these issues as per the individual's opinion. So the values corresponding are grouped as (1)-Not at all, (2)-Several days, (3)-More than half the days and (4)-Nearly every day. We also converted the multi-class problem to a binary problem. Of these, we mapped response 1 and 2 to class 0 and responses 3 and 4 to class 1.



***Learning and Modelling***

We used a test-train split to create train and test data from the survey. We tried out various models for performing classification, including:

- Random Forest (RF)
- Multi-Layer Perceptron (MLP)
- Gradient Boost (GB)
- Adaptive Boost (AB)
- Extreme Gradient Boost (XGB)
- Support Vector Classifier (SVC)

- Categorical Boost (CatBoost)

After data-preprocessing, we did K fold cross-validation for finding out best-performing models. We also did GridSearchCV for binary classification problems to find best hyper-parameters for each of the best performing models. Of these models, RF, MLP, GB/XGB, and AB performed better than the rest.

***Design Choices***

Based on the target variables in the problem and the kind of models that are inherently multiclass, ensemble methods were the first choice. To start with, our first base classifier was Random Forest. But as the performance in terms of accuracy scores was not good enough, other models like K nearest neighbours and Bernoulli Naive Bayes were also implemented. The results were not promising, even after tweaking the hyperparameters. The prediction seemed to be weak and the accuracy had saturated. Thus, the next step was to use gradient boosting methods. Those gave slightly better results but were computationally tiring. We also tried to train a Multi-Layer Perceptron classifier but results were mediocre. Being limited with no grid search for these models due to the target being non-binary, there was very little that could be done for this data. The last step was to convert the target to binary class variables and use grid search, which gave around 80% accuracy with the best performing models from the multiclass problem.

***Results***

In data pre-processing, we tried both dropping the missing values and imputing them. Imputation of the missing data resulted in an increase of accuracy in comparison to just dropping the data.

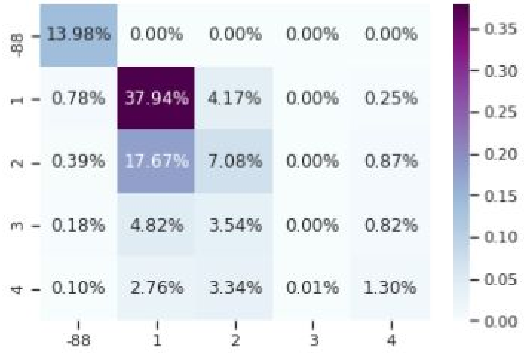
Dropping:

| Down | Anxious | Worry | Interest |
|---|---|---|---|
| XGB = 0.5764<br>AB = 0.5731<br>MLP = 0.5712<br>RF = 0.5264 | XGB = 0.4753<br>MLP = 0.4718<br>AB = 0.4656<br>RF = 0.4117 | XGB = 0.5435<br>AB = 0.5415<br>MLP = 0.5411<br>RF = 0.4851 | XGB = 0.5687<br>AB = 0.5648<br>MLP = 0.5595<br>RF = 0.5050 |

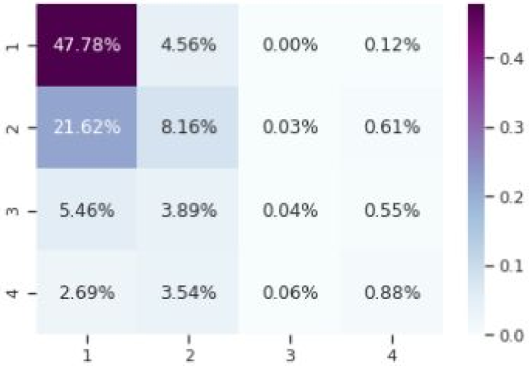
Imputation:

| Down | Anxious | Worry | Interest |
|---|---|---|---|
| XGB = 0.6191<br>MLP = 0.6141<br>RF = 0.5773<br>AB = 0.5505 | XGB = 0.5472<br>MLP = 0.5187<br>RF = 0.5019<br>AB = 0.4794 | XGB = 0.5884<br>MLP = 0.5846<br>RF = 0.5500<br>AB = 0.2501 | XGB = 0.6029<br>MLP = 0.5966<br>RF = 0.5607<br>AB = 0.5535 |

| | | | |
|---|---|---|---|
| | | | |



Dropping

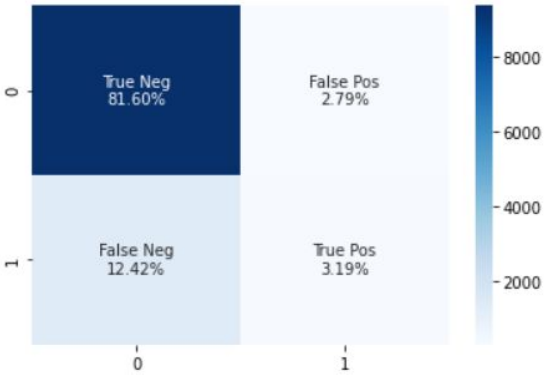


Imputation

Each of the target variables performed differently in terms of accuracy.
“Down” as a target variable gives us the best performance with Gradient Boosting classifier.

Binary Classification:

| Down | Anxious | Worry | Interest |
|---|---|---|---|
| MLP = 0.8462<br>GB = 0.8504<br>RF = 0.8401<br>AB = 0.8503 | MLP = 0.7022<br>GB = 0.7495<br>RF = 0.7254<br>AB = 0.7515 | MLP = 0.7892<br>GB = 0.8026<br>RF = 0.7920<br>AB = 0.8050 | MLP = 0.8324<br>GB = 0.8343<br>RF = 0.8218<br>AB = 0.8353 |



*Binary Classification of Down target variable*

***Key Insights***

The dataset was unique in terms of how it performed with less or more data. When the rows with missing data were dropped to improve upon data quality, it was seen that the classification results were not promising. On the other hand, if raw data with no pre-processing or dataset with imputed values were used, the results improved in both multi-class and binary cases. The strikingly odd observation was that when raw and imputed dataset results were compared, accuracy scores were found to be quite similar for most of the classifiers. The single exception was Random Forest, which performed better on imputation but was inferior in prediction, thus, had little overall improvement. Furthermore, when using feature selection on the data, only MLP classifiers had some benefits - that, too, only in a specific combination with "Interest" as target. Amount of data was also the dominating factor when datasets with top features were compared.

### ***Conclusions***

The stated objectives of this project were to gain insights about the COVID-19 pandemic and its effects with machine learning techniques on big data sources. Each of the tasks in this project tackled the pandemic from a different angle, with different methods used for analysis. At the end of the analysis, we had results showing visually discernible effects of lockdowns and mask mandates, a comparison of the standard and learning-based time-series projection modeling, and classification analysis results which strongly indicates that the state of mental well-being can be identified from other features. While COVID has and will keep affecting different aspects of our daily lives, we also have the ability to change the course of the pandemic with strict mandates. Otherwise, the upward trends in projections will be the dire reality in the months to come.

### *References*

[1] COVID-19 Map. Johns Hopkins Coronavirus Resource Center. (2020). https://coronavirus.jhu.edu/map.html

[2] Robert, A. (2020). Lessons from New Zealand's COVID-19 outbreak response. The Lancet Public Health. https://doi.org/10.1016/s2468-2667(20)30237-1

[3] The key to Viet Nam's successful COVID-19 response: A UN Resident Coordinator blog. UN News. (2020). https://news.un.org/en/story/2020/08/1070852

[4] COVID-19 Secondary data and statistics. CDC.gov. (2020). https://www.cdc.gov/library/researchguides/2019novelcoronavirus/datastatistics.html

[5] Best, R., & Boice, J. (2020). Where The Latest COVID-19 Models Think We're Headed - And Why They Disagree. FiveThirtyEight. https://projects.fivethirtyeight.com/covid-forecasts/

[6] COVID-19 Projections Using Machine Learning. (2020). https://covid19-projections.com/

[7] An, C., Lim, H., Kim, D., Chang, J., Choi, Y., & Kim, S. (2020). Machine learning prediction for mortality of patients diagnosed with COVID-19: A nationwide Korean cohort study. https://www.nature.com/articles/s41598-020-75767-2

[8] Schmitt, M. (2020, November 23). How to fight COVID-19 with machine learning: Data Revenue. https://www.datarevenue.com/en-blog/machine-learning-covid-19

[9] Alonso, S.G., de la Torre-Díez, I., Hamrioui, S. *et al.* Data Mining Algorithms and Techniques in Mental Health: A Systematic Review. *J Med Syst* 42, 161 (2018). https://doi.org/10.1007/s10916-018-1018-2

[10] Ćosić K, Popović S, Šarlija M, Kesedžić I, Jovanovic T. Artificial intelligence in prediction of mental health disorders induced by the COVID-19 pandemic among health care workers. Croat Med J. 2020;61(3):279-288. doi:10.3325/cmj.2020.61.279

[11] Omar Al Omari, et al., "Prevalence and Predictors of Depression, Anxiety, and Stress among Youth at the Time of COVID-19: An Online Cross-Sectional Multicountry Study", Depression Research and Treatment, vol. 2020, Article ID 8887727, 9 pages, 2020. https://doi.org/10.1155/2020/8887727